DOCUMENT RESUME

ED 349 319 TM 018 923

AUTHOR Anderson, Paul S.; Hyers, Albert D.

TITLE Quantitative Comparisons of Difficulty, Discrimination and Reliability of Machine-Scored Completion Items and Tests (in the MDT Un-Cued Answer-Bank Format) in Contrast with Statistics from Comparable Multiple Choice Questions: The First Round of Results.

PUB DATE Apr 91

NOTE 29p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 4-6, 1991).

PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *College Students; Comparative Testing; *Computer Assisted Testing; Correlation; *Difficulty Level; Higher Education; *Multiple Choice Tests; Scoring; Statistical Significance; *Test Reliability; Test Validity

IDENTIFIERS Answer Banks; *Item Discrimination (Tests); *Multi Digit Tests

ABSTRACT

Three descriptive statistics (difficulty, discrimination, and reliability) of multiple-choice (MC) test items were compared to those of a new (1980s) format of machine-scored questions. The new method, answer-bank multi-digit testing (MDT), uses alphabetized lists of up to 1,000 alternatives and approximates the completion style of assessment items. Five data sets were analyzed with a total of over 500 postsecondary students. Each student answered between 13 and 30 pairs (total of 110 pairs) of identical questions in both MC and MDT formats. Quantitative measures, including correlation and significance of difference in means, reveal that in comparison to MC questions, the MDT items predictably are: (1) more difficult; (2) consistently better discriminators (higher validity); and (3) continually better providers of test reliability. The evidence firmly indicates that the MDT answer bank format of responses is superior in these quantitative measures to the MC format for questions where either could be used appropriately. The MDT format could be widely used in education to improve measurement without raising costs. One table and four appendixes are included. Appendixes 1 and 2 provide background material and a description of the data sets. Appendixes 3 and 4 contain 15 tables of statistics describing the data sets. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
PAUL S. ANDERSON
ALBERT D. HYERS
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

QUANTITATIVE COMPARISONS OF DIFFICULTY, DISCRIMINATION AND RELIABILITY OF MACHINE-SCORED COMPLETION ITEMS AND TESTS (IN THE MDT UN-CUED ANSWER-BANK FORMAT) IN CONTRAST WITH STATISTICS FROM COMPARABLE MULTIPLE CHOICE QUESTIONS: THE FIRST ROUND OF RESULTS

ED 340 819

Paul S. Anderson, Ph.D.
Associate Professor of Geography
Illinois State University
Normal, IL 61761
Vice President, Multi-Digit
Technologies (MDT) Corp.
107 Broadway, Normal, IL 61761

Albert D. Hyers, Ph.D.
Assistant Professor of Geography
North Adams State College
North Adams, MA 01247

TM018923

BEST COPY AVAILABLE

Quantitative Comparisons Of Difficulty, Discrimination And Reliability Of Machine-Scored Completion Items And Tests (In The MDT Un-Cued Answer-Bank Format) In Contrast With Statistics From Comparable Multiple Choice Questions: The First Round of Results

(Final Version as submitted to ERIC.)

Paul S. Anderson, Ph.D.
Associate Professor of Geography
Illinois State University, Normal, IL 61761
Vice President, Multi-Digit Technologies (MDT) Corp.
107 Broadway, Normal, IL 61761

Albert D. Hyers, Ph.D.
Assistant Professor of Geography
North Adams State College
North Adams, MA 01247

ABSTRACT

We compare three descriptive statistics (difficulty, discrimination, and reliability) of multiple choice (MC) items to those of a new (1980s) format of machine-scored questions. Called answer bank multi-digit testing (MDT), the new method uses alphabetized lists of up to 1000 alternatives and approximates the completion style of assessment items. We analyze five data sets with a total of over 500 post-secondary students. Each student answered between thirteen and thirty pairs (total of 110 pairs) of identical questions in both MC and MDT formats. Quantitative measures, including correlation and significance of difference in means, reveal that in comparison to MC questions, the MDT items are predictably 1) more difficult, 2) consistently better discriminators (higher validity), and 3) continually better providers of test reliability (KR-20). The evidence firmly indicates that the MDT answer bank format of responses is superior in these quantitative measures to the MC format for questions where either method could be used appropriately. The MDT format could be widely used in education to improve measurement without raising costs.

[A paper presented to the 1991 Annual Meeting of the National Council on Measurement in Education, Chicago, Illinois, April 4-6, 1991.]

I. INTRODUCTION AND OBJECTIVES OF INQUIRY

This paper is the first to report a series of quantitative comparisons between multiple choice (MC) questions and an alternative format that can also be scored on the same currently available Optical Mark Reader (OMR) equipment. The alternative format, developed in the 1980s, has been called various names that, in combination, describe its main characteristics: long-menu, long-list, key-list, answer-bank, un-cued, multi-digit testing (MDT). We refer to it here as the MDT answer bank format and discuss it below in Section II.A. and in Appendix 1.

The objectives of our research are to collect and analyze sets of comparable data of paired test items in the MDT and MC formats in order to understand the differences and similarities between them. In this paper we present our initial results of quantitative measurements of these two (MC and MDT) machine-scored methods of educational measurement. We focus primarily on statistics of difficulty, discrimination, and reliability. By definition and design, this study is exploratory. It leads us to suggest avenues for additional analyses in the conclusion.

TM018923

II. METHODS

A. Methodology for Obtaining Data

The basic methodology for obtaining data is to ask students identical questions in the two item formats being compared. The multiple choice (MC) format used in this study is the traditional approach. Four or five choices, sometimes including "none of the above," are offered after each item stem.

In contrast, the MDT answer bank method uses a separate **alphabetized** and numbered long list of up to 1000 alternatives to serve as a "bank of answers" for many questions, even for several examinations in a semester. Beside each entry on the list there is a multiple-digit code number (usually three digits) to mark on a specially designed answer sheet that can be scanned with standard optical mark readers. (See Appendix 1 for an explanation.) When no correct answer is on the answer-bank list, the student is instructed to write the desired response in a blank area for rapid manual scoring on the answer sheet and also mark for machine-scoring the MDT code number for the first letter of the key word of the desired response. One fundamental aspect of such a long list is that the students are neither expected nor encouraged to read the entire list to find or recognize the correct answer. Instead, the student is to decide upon the response and then simply go directly to that response on the alphabetized list. Based on many dozens of examinations with thousands of students plus other research (such as Anderson (1988) on student attitudes about styles of questions), we can say with assurance that this format is very nearly equivalent to completion (fill-in-the-blank) style questions, but with the advantages of machine scoring.

> [The MDT answer bank format has been described since 1983 in numerous papers and publications by its principal developer, Paul S. Anderson, one of the co-authors of this paper. See Bibliography. Especially the 1987 book and the 1990 AERA/NCME paper (both available on ERIC microfiche) are good explanations of the MDT method. The brief explanation in Appendix 1 also discusses the unique features of reconsiderative scoring made possible by the MDT answer bank method and software.]

B. Descriptions of Data Sets

We collected four of the five data sets in our courses at one large public Mid-Western university and at one small public New England college. The course subjects are different; one data set (PG90) is from a physical science course (physical geography) and three data sets (WG84, WG87, and WG90) are from a social science course (world regional geography). The fifth data set (MED89) is from an examination in the field of medicine.

Two data sets (WG84 and WG87) use two matched groups of students who answered the questions in only one format. The other three data sets are from students who answered the MDT questions first, after which, in the examination period, they answered the MC versions of the same questions, without being allowed to re-examine the MDT version. The data sets are from actual in-class examinations and, therefore, are subject to the advantages and limitations of such data. In a few cases some of the responses did not count toward the course grade, but the researchers have taken great care to be fair to the students while also motivating them to participate to the best of their ability.

Appendix 2 presents background information about these five data sets and Table 2 below summarizes that information in the same order as most of the tables and discussions later in this paper. The data sets are not error-free, are not controlled experiments, and are not presented as representing anything other than themselves and equivalent populations. However, consistent statistical results from five data sets with a total of 110 pairs of identical items could constitute empirical evidence of meaningful differences and similarities between the two formats of questions. As stated earlier, this study is exploratory.

Table 1: Summary of Sizes of Five Data Sets

	Data Set Name	Number of Examinees	Number of Items (Pairs)	Number of Responses			
1	WG84	77 MDT 82 MC	20	1540 MDT 1640 MC			
2	WG87	77 MDT 84 MC	30	2310 MDT 2520 MC			
3	WG90	160 MDT + MC	20	3200 MDT 3200 MC			
4	WGALL (Rows 1+2+3)		70	7050 MDT 7360 MC			
5	PG90	42 MDT + MC	13	546 MDT 546 MC			
6	MED89	93 MDT + MC	27	2511 MDT 2511 MC			
7	TOTAL		110	10107 MDT 10417 MC			

C. Criteria for Statistical Analyses

We process and analyze the collected student responses to determine how MDT and MC versions of questions differ, if at all, in terms of difficulty, discrimination, and reliability. The software that scores the tests also provides these basic numerical indices.

Difficulty: The index of difficulty is the decimal expression of the proportion of students answering a question correctly. For example, values of 1.000 and 0.500 indicate, respectively, that all (100%) and half (50%) answered correctly.

Discrimination: The index of discrimination used with three of the five data sets is the decimal expression of the following calculation for each item in each test. Discrimination values can range from -1.000 to 0.000 to +1.000 for each item.

$$\frac{\text{\# of students correct in highest 27\%}}{\text{total \# of students in highest 27\%}} \quad \text{minus} \quad \frac{\text{\# correct in lowest 27\%}}{\text{total \# in lowest 27\%}}$$

Discrimination values for data set WG84 were not readily available to us. The data set MED89 for medical questions used the point-biserial correlation coefficient. We anticipate rescanning and scoring the WG84 answer forms to obtain discrimination values and also doing comparative work using the point-biserial statistics.

Reliability: The index of reliability used is the Kuder-Richardson formula 20 (KR-20). Although higher values (such as 0.80+) are desirable, KR-20 values are strongly influenced by the number of items in the examination or subgroup of questions. Therefore, comparisons between matched subtests (that is, those with paired MC and MDT versions of the questions) are appropriate, but comparisons between data sets require greater care in interpretation.

Analytical Statistics: We used a statistical analysis package to derive the correlations, regressions, tests of statistical significance, and other values reported. For appropriate normally distributed data, we used parametric statistics (Pearson's r, t-test, and analysis of variance). When data were not normally distributed or did not otherwise meet the necessary assumptions, we were sometimes able to transform them to meet the requirements for parametric tests. Otherwise, we used equivalent non-parametric procedures (Spearman's rank

order correlation, Wilcoxon signed ranks, and Kruskal-Wallis tests). Nearly all non-normal distributions were skewed to the left.

P-Values: We have set the following probability levels to assess the likelihood that the observed differences would not be merely the result of error in sampling.

> $P < 0.0100$: Strong indication that the observed difference is NOT the result of chance.
>
> $0.0100 < P < 0.0500$: Reasonable probability that the observed difference is not the result of chance.
>
> $0.0500 < P < 0.1000$: Modest security in having confidence that the observed difference is real and not the result of the sampling procedure. We can say that a trend or tendency is observed. An observed trend has greater importance if similar analyses with larger sample sizes yield reasonable or strong probabilities concerning the same issue being examined.
>
> $P > 0.1000$: Although a difference may be observed, the likelihood of that difference being the result of the sample (and not of the issue) is too great to merit meaningful interpretations.

III. TABULATED STATISTICAL RESULTS

Summaries of the data are presented in the discussion sections of the paper. The full statistical results are in the tables of Appendices 3 and 4. Appendix 3 presents the base data by item; Appendix 4 contains the calculated statistics for each data set. Both appendices have the same internal order by data set: WG84, WG87, WG90, PG90, and MED89.

Appendix 3: The base data for difficulty levels and discrimination values for identical questions in the two different formats are in Appendix 3. For reference purposes, these data on a per-item basis are important statistics derived from the individual student responses. However, they are primarily an intermediate step toward the desired analyses. Summary values from these data are presented in the tables of Appendix 4.

Appendix 4: Key calculated statistics (such as means, differences, correlations, and significance levels) for difficulty, discrimination, and reliability are incorporated into the tables in Appendix 4. Within each table the columns are lettered and the rows are numbered. In this way we can identify each cell of data. For example, the mean difficulty index for the MDT questions from data set WG87 is in Appendix 4, Table 2, Cell B5 (abbreviated as AP4-T2-B5).

We present our results below under the headings of difficulty, discrimination, and reliability. These measures overlap and can influence each other.

IV. DIFFICULTY

A. Differences in Difficulty Levels

Item difficulty has major influence in education. Student grades, including passing or failing a course or an entire year of schooling, are strongly influenced by the level of difficulty of tests. Although administrative policies have influence, teachers are in ultimate control of the difficulty levels of their courses, tests and items. Furthermore, as we shall discuss later, item difficulty does influence other statistical values for discrimination and reliability. Therefore, we desire to understand the relationship of MDT format questions in relation to the currently popular "standard" MC format.

MDT answer bank questions are more difficult than MC questions for otherwise equivalent items. The greater the number of alternative answers from which a student can choose a response, the greater is the difficulty and the lower is the difficulty index. The answer bank method utilizes this fact by providing up to 1000

alternative answers. We recommend lists with well over 500 entries for post-secondary examinations. A lengthy list is a major feature and fundamental principle of the MDT method. However, the importance of list length is poorly understood by uninformed users who sometimes use short lists and then wonder why their results are different. **The length of the answer bank intentionally attempts to approach the characteristics of the fill-in-the-blank (FIB) item format where the "list" is theoretically infinitely long.** All five sets of the data confirm that the MDT method is more difficult by an average of 10% to 12%, with variations as noted below. For four of the five sets, the differences are statistically significant (at the $P < 0.01$ level or stronger), as is the case with many of the subsets with smaller sample sizes.

The "Difference in Mean Difficulty" levels (DMD) for the 110 paired items was 9.9 percent lower for MDT questions than for MC items (DMD = -0.099 with $P < 0.0000$). If we consider that the base difficulty of the MC questions was 71.3 percent, then the difficulty of the MDT items was a 13.9 percent reduction of the student scores (-13.9% = -0.099/0.713). We refer to this as the "Proportionate Change in Difficulty" or "PCD", in which the MC difficulty index serves as the base number. Regardless of whether the DMD or PCD value is used, the result is clear. With other variables held constant or controlled, student scores on MDT questions are lower because the questions are more difficult than matched MC items, not because of student abilities, efforts or learning.

The differences of mean difficulty (DMD) for the five data sets range from -0.040 (WG87) to -0.190 (PG90). This nearly five-fold spread is primarily the result of the proportion of the questions that were easy or difficult, a topic that we now consider.

B. Subsets Divided by Difficulty levels

When a large majority of examinees know or can derive a correct answer, that item is easy and has a high numerical index of difficulty (e.g., greater than 0.80 or 0.90). Because the factors of recognition and guessing are minimal when nearly everyone knows the answer without even reading the five choices, the difference in difficulty between identical MC and MDT questions is minimal, in accordance with the laws of probability. When a question is very difficult and the five choices are good enough to minimize the process of elimination, then the recognition and guess factors become important. Therefore, in accordance with the laws of probability, the 4-choice and 5-choice MC questions should yield difficulty indices of approximately 0.25 and 0.20, respectively, while MDT P-Values should approach 0.002 and 0.001 for answer banks with 500 to 1000 responses. We test this hypothesis in two ways. First we compare subsets divided by difficulty levels. Second, we calculate correlations and regressions for the paired data.

The data (see Appendix 4) consistently show that easy questions have less differences in difficulty between the two item formats than do difficult questions. Note that we have used several different cut-off levels to create subsets for "easy," such as > 0.800 and > 0.750 and > 0.667. We have done so with several of the data sets in our exploratory efforts because for some statistical analyses, the numbers of items in the subsets were too small to be meaningful. We do not conclude that one cut-off level is better than another.

For the total sample of 110 pairs of items, the subsets of hard, intermediate, and easy (also called difficult, medium, and facile) questions are divided at MDT difficulty levels of 0.333 and 0.667. We find in the facile subset that the difference of the mean difficulty (DMD) levels is only -0.034 (i.e., DiffMDT - DiffMC), but the difference is -0.155 and -0.150 for medium and hard, respectively (all with $P < 0.01$, our most rigorous level of statistical significance).

Equally revealing is the amount of the proportionate change in difficulty. The PCD for easy questions is only -4.0% (-0.034/0.854), while the intermediate level questions have a change of -23.0%. Even more noteworthy is the PCD = -44.2% for the hard questions.

C. Correlations and Regressions of MDT and MC Difficulty Levels

For the combined (total) data set with 110 pairs of items from the five data sets, the correlation of MDT and MC difficulty levels is $r = 0.8501$ ($P < 0.0000$). This means that 72.3% (r^2) of the variance in MC is explained by the MDT difficulty index. Of all of the statistically significant correlations, the highest was $r = 0.9783$ (for WG90 with 10 questions). We found lower coefficients in the subsets divided by MDT difficulty levels. For example, r equals 0.4386 for the 46 intermediate difficulty items of the combined data set. But those lower results are in large part caused by limiting the X-axis of MDT difficulty levels to only the range from 0.333 to 0.667. The tables in Appendix 4 contain these various results, including those that are not statistically significant, so that each reader can consider the impact of difficulty range and also sample size.

The Y-axis intercept for the linear regression line of the combined data set (110 pairs of items) is 0.2547 with a standard error of 0.0293. The slope of the regression line in this combined data set is 0.7470 with standard error of 0.0445. The intercept and slope combine to say that at least for the total data set of 110 pairs of items, the data results are extremely close to what the laws of probability say about these two formats of questions: With 1000 possible responses in an answer bank, there is a negligible 0.1% probability of guessing correctly, while a correctly constructed five-choice multiple choice item yields 20% correct responses by random guessing.

D. Comments on the Findings Concerning Difficulty Levels

Given equivalent questions that could be asked appropriately in either the MC or MDT formats, increased academic rigor is attained with the MDT answer bank method. The increase in rigor is substantially greater for medium and hard questions than for easy questions. Although the results presented above cannot be generalized to apply to other populations of students and test items, the indication is that the greater difficulty of the MDT answer bank format is a meaningful (not trivial) amount in terms of student scores on tests.

In light of the research findings, teachers who use the MDT answer bank format should adjust their grading systems in one of several ways listed below. Such adjustments should be in proportion to the amount of a total test that has been appropriately converted into the MDT format from the MC format.

1) Use a curve with predetermined percentages for each grade bracket,
2) Lower the predetermined cut-off levels for each grade bracket, or
3) Acknowledge to themselves and to their students that the course expectations have been raised by the use of the MDT method.

From our experiences in college-level classrooms, both of us have independently adopted levels of 80% to 85% as the bottom of the "A" bracket, with "C" grades commonly allowed as low as the mid-50s. Many students initially have difficulty understanding that machine-scored tests do not automatically use the 90, 80, 70 cut-offs for A, B, and C grades, respectively. The ingrained idea that "90 and above is an A" is a prevalent myth in American education, almost as if it were the Eleventh Commandment. Perhaps the MDT answer bank method can assist to change this inappropriate tradition.

V. DISCRIMINATION

A. Face Validity and Content Validity

Although item difficulty is important, the validity of each item has even greater meaning. However, validity has several distinct aspects, each of which is either more subjective or harder to measure and interpret than difficulty.

Face validity is essentially the appearance or perception of an assessment method being a valid measurement of student ability. The face validity of the MDT answer bank method is largely tied to the notion that this method is essentially a machine-scored version of completion or fill-in-the-blank (FIB) questions. We do not have quantitative evidence about face validity, but we note that earlier studies have made relevant

comments. Veloski et.al. (1988, abstract) says this method "... provides important advantages that strengthen its face validity. The Un-Q [MDT] format can be used to test either simple recall or certain higher level problem-solving skills that cannot be tested by MCQ."

Anderson (1988) reports that student attitudes about the MDT method are quite similar to their attitudes about the FIB method, both in preference for the various test formats and in student perception of the ability of each format to evaluate students In other words, students who have experienced the answer-bank method do perceive a greater validity than what is offered by traditional multiple choice questions.

We are two of the most experienced users of the MDT method. Our joint opinion is that the face validity of the MDT answer bank method is at least equal to and is probably better than the MC method. However, we are not surprised when teachers who do not fully understand the method have doubts, especially when they do not appreciate the great importance of having a very long list appropriate to the grade level of the students. Face validity is a rather subjective matter, especially from people who have not used the method.

Content validity is commonly assessed by educators who are "subject masters" for the topics of the assessment. We are both professional geographers employed for a combined total of thirty years in university education. We are not perfect at writing test items, but we are certain that other geographers would find the content of our questions used in this research to be appropriate for geography classrooms, whether in the MC or MDT formats. Similarly, the items used in the MED90 data set are also the product of subject masters, in this case, in family practice medicine.

B. Construct Validity (Item Discrimination)

Construct validity, which we refer to here in the context of educational measurement as item discrimination, refers to how effective an assessment item is at distinguishing between students in accordance with their ability. It is a more quantified form of validity, but it is hard to measure. The "index of discrimination" measurement used in three of the five data sets (see Section II.C. above) contrasts the top 27% with the lowest 27% of the students, AS DETERMINED IN THE SPECIFIC TEST BEING ANALYZED. Therefore, the discrimination calculation as used here is NOT based on other judgements of ability, such as scores on standardized tests.

The mean discrimination values for the 63 pairs of items in three combined geography data sets (WG87, WG90, PG90) is the largest sample available to us. We named this combined data set GEOG3. The tables in Appendix 3 reveal that only twice out of 126 calculations of discrimination were the values negative, once for MC (-0.167 in PG90) and once for MDT (-0.095 in WG87). The MDT versions yielded higher discrimination indices in 43 of the 63 pairs of items (68.3%). The maximum and mean discriminations, respectively, were 0.711 and 0.349 for MC versions and 0.833 and 0.399 for MDT items. As teachers, we would prefer that the items on our tests have higher discrimination values. As researchers, we note that the MDT answer bank method yields higher mean discrimination values than does the MC method with the 63 items in this combined data set. The statistically significant absolute difference of 0.050 between the means and the 14.3% relative increase (0.050/0.349) over the MC mean are in the direction we expected. Although almost any improvement in item discrimination should be beneficial to education, we investigate influences of difficulty before commenting on whether the increase has meaningful importance for educational practices.

C. The Influence of Difficulty on Discrimination

The statistic called index of discrimination is strongly influenced by item difficulty. Questions with very high or very low difficulty indices cannot attain high values in item discrimination because the formula is based on proportions of respondents answering the item correctly or incorrectly. In the combined data set GEOG3, the means were suppressed by the large number of very easy items (12 items with MDT difficulty greater than 0.90) plus 5 difficult items with MDT difficulty less than 0.33. The mean discrimination indices for those 12 very easy items are essentially identical (0.135 and 0.137) for the MDT and MC versions of the questions. These values are quite low and raise the issue of whether or not very easy questions should be included on examinations of either format of questions. Issues of morale building for weak students and criterion referenced assessment

are possible justifications for including a few such very easy items. But easy items are not very good discriminators and they cloud the current comparison of the MDT and MC formats.

The 5 difficult items have mean difficulties of 0.201 (MDT) and 0.304 (MC), with the difference resulting largely from one MC item (#3 in PG90) with a value of 0.630. The mean discriminations are 0.114 and 0.365 for MDT and MC, respectively, with a difference of 0.251 in favor of the multiple choice method. Although such a large difference is noteworthy, we hypothesize that this is a reverberation from the limited selection of five choices in MC items. When a question is truly difficult, students use the "process of elimination" to improve their chances for a correct answer. The more capable students (more prepared or more intelligent?) can eliminate better, and the net result is better discrimination. We raise the issue of whether or not this "good guesser" aspect of assessment with MC questions is desirable. In light of the current debate in the USA about the pros and cons of multiple choice questions, this issue merits further research and debate. For this current study, our conclusion is that the difficult questions cloud the real concern as to which, if either, format provides better discrimination. Item discrimination is appreciably influenced by item difficulty, especially near the extremes.

In order to focus on the central or "intermediate difficulty" questions, we reprocessed a subsample of the combined data set GEOG3 using two different cut-offs of MDT difficulty. For cutoffs at 0.33 and 0.90 (see Appendix 4, Table 10, Column B), the results show mean difficulty levels of 0.591 and 0.698 and mean discrimination indices of 0.500 and 0.402 for the middle subgroup of MDT and MC questions, respectively. For cutoffs at 0.33 and 0.67 (see Appendix 4, Table 9, Column D), mean difficulty levels of 0.507 and 0.650 correspond to discrimination levels of 0.528 and 0.419. These difficulty and discrimination values are more in keeping with our objectives for educational assessment, that is, to have a meaningful system to inform students of their level of learning and to allocate grades.

D. Interpretation

The differences between the mean discrimination indices (0.098 and 0.109) for the two data sets with intermediate difficulty (as described in the previous paragraph) are double the difference for the entire data set GEOG3 (see Appendix 4, Table 9, Column B, line 16). That difference is a 28.1% to 31.2% increase over the base mean discrimination of the traditional MC version of the questions. In other words, these initial analyses show that when item difficulty is controlled for using only items with intermediate difficulty, the MDT answer bank format of questions can cause an increase of more than twenty-five percent (25%) of the item difficulty of equivalent questions in the traditional multiple choice format.

We consider this improvement in item discrimination to be highly noteworthy and truly meaningful for the improvement of educational assessments, especially those tests and exercises that can be machine scored.

VI. ADDITIONAL ISSUES AND ANALYSES

A. Correlations and Regressions of MDT and MC Discrimination Indices

As we see in Appendix 4, statistically significant correlations exist between the discrimination indices of MDT and MC questions. We interpret those analyses in a later paper.

B. Reliability

Using the KR-20 statistic for test reliability, we find that subtests of MDT questions generally yield higher values than do subtests of MC items. Because the KR-20 value is strongly influenced by the sample size and by the difficulty of each item, our future analyses for detailed comparisons will be based upon subtests that are equated as much as possible. The multi-digit response format does require a moderate amount of additional time for each question, in comparison with five-choice questions. In a future paper we will examine the issue of time verses number-of-questions in the context of the test reliability and other issues.

C. Correlations with "Ability"

In 1988, Veloski et al. wrote that the MDT answer bank method (which they called "Un-cued") might "improve identification of marginal examinees." We have briefly examined correlations of student responses/scores with variables that indicate "ability," as defined by course grades, GPA, and scores on standardized tests (ACT and SAT). Our analyses are not yet complete, but early results indicate correlations that confirm Veloski's statement.

VII. Educational Importance of the Study:

Any new method of educational measurement must undergo rigorous scrutiny, including quantitative analyses that reveal its strengths and weaknesses. Those who have used the MDT answer bank method have generally been impressed with its capabilities. Because little prior quantitative work has been done, this paper should have major significance to specialists in educational measurement and to educators at large.

The most important single finding is that item discrimination can be meaningfully increased (by over 25% for intermediate difficulty questions) when the test format is changed from the multiple choice format to the MDT answer bank format. We believe that such results should merit considerable research interests.

The MDT method does not increase the cost of machine scored assessment. It is clearly applicable to almost all varieties of tests, including norm referenced, criterion referenced, teacher-made, and most of the computer-administered systems for both instruction and measurement. This paper contributes to the objective evaluation of this answer bank method.

BIBLIOGRAPHY OF CITED WORKS:

Anderson, Paul S. and Eileen M. Kanzler (1985). "Comparison of Cognitive Achievement in Objective Testing: Multi-Digit and Multiple Choice Tests." Paper presented at the conference of the American Educational Research Association (AERA), Chicago, Illinois, April 4, 1985. [ERIC microfiche ED 260 131]

Anderson, Paul S. (1987). The MDT Innovation: Machine Scoring of Fill-in-the-Blank Tests. Multi-Digit Technologies Corporation, Normal, Illinois. (198 pages). [ERIC ED 307 287 - three fiche]

Anderson, Paul S. (1988) "An Educology of Testing: American Student Attitudes about Test Formats, with Special Reference to the MDT Multi-Digit Testing Method." International Journal of Educology, Vol. 2, Sydney, Australia, pp. 143-184. 1988. [ERIC EJ 394 496]

Anderson, Paul S. (1990). "Initial Experiences with Reconsiderative Scoring." Paper presented at the conference of the National Council for Measurement in Education (NCME), Boston, April, 1990. [ERIC 318 788]

Brailovsky, C. A., G. Bordage, T. Allen and H. Dumont (1988). "Writing vs. Coding Diagnostic Impressions in an Examination: Short-Answer vs. Long-Menu Responses." Research in Medical Education (RIME) Proceedings, the 27th Annual Conference of the Association of American Medical Colleges, Chicago, Illinois, 11-17 November 1988, pages 201-206.

Veloski, J. Jon, Howard K. Rabinowitz, and Mary R. Robeson (1988). "Cueing in Multiple choice Questions: A Reliable, Valid and Economical Solution." Paper (with abstract) presented to the conference on Research in Medical Education at the meeting of the Association of American Medical Colleges, Chicago, Illinois, November 1988. The paper without the abstract is published in Research in Medical Education (RIME) Proceedings, the 27th Annual Conference of the Association of American Medical Colleges, Chicago, Illinois, 11-17 November 1988, pages 195-200.

APPENDIX 1: BACKGROUND TO MDT MULTI-DIGIT TESTING

The innovation for machine scoring of fill-in-the-blank questions that Anderson first used in early 1983 is described in his book, The MDT Innovation (1987, available in all libraries with ERIC microfiche). The 200-page book discusses the method's origins, initial usage, applications and educational implications, including financial savings and relevance to higher order learning. The MDT innovation is very straight-forward.

A. Completion-style questions are asked in a wide variety of ways that require a discreet word (concept) or numeric response.

B. With an answer in mind, the student locates his/her desired response on an alphabetized long-list "answer bank" appropriate for the subject area. The list is quite important, but it can be made easily by any teacher or shared by teachers of similar courses.

C. The three-digit code number of the response is marked on an answer form that is read by standard optical mark readers of sheets or cards.

D. Microcomputer software processes the student responses, issues their scores and prints distinctly useful reports.

With such a simple and natural methodology, was nothing like this ever tried previously? Chapter 8 in The MDT Innovation (Anderson, 1987) reviews prior efforts fortunately not known to Anderson when he began. Anderson's work is distinctive in two crucial ways.

First, no previous researcher had significantly gone beyond one hundred responses, that is, responses with two digits. Anderson began his work with three-digit numbers that allow up to 1000 terms and concepts in the answer bank. Each list is intentionally long to nearly eliminate the "I'll recognize the answer if I see it" effect. Short lists encourage and permit searching that waste time when students try to recognize answers. This is a fundamental difference from "matching" tests. List length should be in consideration of the grade level of the students, but totally at the discretion of the test writer. For university students, Anderson recommends lists of over 500 terms. Furthermore, he uses response number 999 to mean "no reasonable answer is on this list." When responding with 999, the student also writes the correct word response in a blank space on the answer sheet. The evident strength of the MDT method is the high degree to which it emulates completion-style (fill-in-the-blank) responses, of which the academic rigor and validity are well established.

Second, earlier researchers did not develop any computer program to do the scoring. Consequently, Anderson was the first to observe and use the tremendous time-saving advantages of these computer-scored completion tests. Even more important, he recognized and made operational the capability to generate some truly distinctive feedback for teachers and students. The most innovative feedback involves the ability to print the actual word responses of the students and of the teacher's answer key. The MDT responses with printed words in the Individual Student Report give far more assistance to the student than do the multiple choice responses of A B C D or E. The Item Analysis tabulation of responses gives valuable information to the teacher about student learning. Of special importance is the ability of the Item Analysis to print on the screen or paper the actual word (or numeric) responses of the students.

Even though he worked with these reports regularly since 1983, Anderson needed nearly four years before the second innovation became evident to him. Only in the 1989-90 academic year has this second innovation become a reality in a software program called RECON for reconsiderative scoring.

Reconsiderative scoring occurs when a student's response to a multiple choice or completion (fill-in-the-blank) question is read by the teacher prior to the determination of correctness and point value. With software, the above described Item Analysis tabulation presents on the monitor screen the student responses. The teacher simply moves the cursor bar up or down to highlight any fully or partially correct response and then allocate the number of points. By repeating this decision making process, the teacher can designate multiple responses that receive some credit. When the teacher is finished, the computer allocates the designated points to every student's point total. The full power of RECON for reconsiderative scoring is discussed by Anderson (1990).

APPENDIX 2: DESCRIPTIONS OF DATA SETS

Data Set WG84: In the spring semester of 1984, four professors cooperated in a study comparing the impact of MDT, MC and hand-written fill-in-the-blank (FIB) formats on student learning. The published results contain a description of the groups of the students in the data set (Anderson, Hill, Naim and Walters, 1985). The separate sections of the world geography course were treated as equals. In a Latin-Square arrangement, each section took one version (MC, MDT, or FIB) of twenty map/place name questions during each of three mid-term examinations. The statistical comparisons of difficulty, discrimination, and reliability were not a focus of that study. Only the data for difficulty level from Test 1 (with 81 students with the MC version and 77 with the MDT version) are available for this paper.

Data Set WG87: In the fall semester of 1987 a total of 161 students in two "back-to-back" sections of one professor's (PSA) world geography course yielded the data set called WG87 (for World Geography 1987). For all four mid-term tests in the course, the students were randomly divided (using odd/even last digits of ID numbers and the first letters of the last names) into eight groups, being four in each section. The eight groups were combined for analyses according to the test format for each respondent.

As pre-announced, this group of students was given either the multiple choice or the MDT answer bank version of identical questions in each of four 75-minute tests (excluding final exam) during the semester. Students in Groups 2 and 8 had only multiple choice (MC) questions (total of 40 items for 80 points). [Note: All students also answered one manually scored essay question (20 points) on each test, but those results have not been incorporated into this current conference paper.] Groups 1 and 7 had 30 questions on each test in the MDT format, plus 10 MC questions and the essay. The other groups alternated between the two versions of the test, on a pre-announced schedule. This WG87 data set is the results from the 30 items asked in the MDT or MC formats on the fourth test, by which time the students had their maximum familiarity with the MDT method. This data set and WG84 share the characteristic that all students were answering questions that counted toward their course grades, thereby being uniformly motivated to do their best efforts.

Data from the final examination, final course grade, year in school (61% freshmen, 24% sophomores), ACT scores (average 21), gender (42% male), age (median 19), major (51% education and social sciences), GPA (average 2.73/4.0), are also available for analyses. As expected, the randomly selected groups of respondents for each test format were found to be similar to each other and to the entire class in terms of these characteristics.

Data Set WG90: During the fall 1990 semester, 160 students in two sections of a world geography course (by PSA) volunteered to answer up to ten directly comparable MC questions after they completed each of four MDT-format tests. Only three students declined to participate. The data set is the combination of those questions from the third and fourth examinations, with ten pairs of questions from each. The data set is split into WG90T3 and WG90T4 for some of the analyses that look for differences between them.

The above three data sets are all from several classes of one professor's course in world regional geography that he has taught to approximately 2500 students since the spring semester of 1983. The authors also have records on how students in previous and on-going semesters have responded to the same and similar questions. These data are drawn upon when further comparisons are needed.

Data Set PG90: During the fall semester of 1990, another data set was collected by ADH. The respondents were 42 students enrolled in an introductory physical geography course, called here PGCOM. This group had a median age of 20.5, including 11% "nontraditional" students over 24 years old. Freshmen (36%) and sophomores (36%) were the majority in this 62% male group. Business (34%) and social sciences (19%) were the most common majors, but another 30% were undeclared. At the beginning of the semester the mean cumulative GPA of the 42 students was 2.49/4.0 and the average composite score of the 27 who took the SAT was 893.

Students in this PGCOM group participated in data collection involving separately administered, parallel versions of test questions in both the MDT and MC formats. For each of three midterm examinations, students

received extra credit for correctly answering five multiple choice questions that were exact parallels of MDT items in the main body of the test. The students did not see the MC items until after first handing in all other examination materials. Of the fifteen pairs of items, two were removed from the analyses because our suspicion that question wording strongly influenced the responses and the difficulty and discrimination statistics. The thirteen items have been combined into this one data set.

Data Set MED89: We have been given access to statistical data from an independent research study involving 93 medical residents of family practice medicine. Each examinee answered 27 identical pairs of questions in the MDT and MC formats in the fall of 1989. We consider the MED89 data to be of high quality, being obtained from a carefully selected sample of highly motivated medical residents. These were the only graduate level examinee in the five data sets; they constitute the most homogeneous data set. A team of experts in family practice medicine conscientiously selected the twenty-seven pairs of items on the examination, making this the most refined data set. The discrimination calculation of the MED89 data set is the point-biserial correlation. Our use of the MED89 data is to show comparisons with our own data. The principal researcher, J. Jon Veloski, presented other very interesting results from that study at this NCME conference (session Q5, but his name is not on the NCME program).

APPENDIX 3

Table 1: Difficulty and Discrimination Indices for Data Set WG84

Item No.	MDT Diff. Category	MDT Diff. Index	MC Diff. Index	Difference in Diff. Indices
19	1	0.064	0.415	-0.350
4	1	0.240	0.732	-0.492
20	2	0.583	0.829	-0.247
8	2	0.595	0.829	-0.234
15	2	0.600	0.585	0.015
5	2	0.602	0.781	-0.179
12	2	0.609	0.793	-0.184
9	2	0.637	0.695	-0.058
14	3	0.720	0.561	0.159
3	3	0.734	0.915	-0.181
16	3	0.743	0.732	0.011
7	3	0.750	0.878	-0.128
18	3	0.762	0.780	-0.019
17	3	0.800	0.817	-0.017
6	3	0.804	0.927	-0.123
10	3	0.812	0.646	0.166
11	3	0.830	0.927	-0.097
1	3	0.871	0.963	-0.092
13	3	0.875	0.890	-0.015
2	3	0.880	0.902	-0.022
Average		0.676	0.780	-0.104
Median		0.738	0.805	-0.095

Notes:

(1) "MDT Diff. Category" refers to a ranking of items by their MDT Difficulty Indices, with divisions made at 0.333 and 0.667:
1 = Difficult Item
2 = Intermediate Item
3 = Easy Item.

(2) Other abbreviations are as follows: "MDT" = Multi-digit Answer Bank Format; "MC" = Multiple Choice Format; "Diff." = Item Difficulty; and "Discr." = Item Discrimination.

APPENDIX 3

Table 2: Difficulty and Discrimination Indices
for Data Set WG87

Item No.	MDT Diff. Category	MDT Diff. Index	MC Diff. Index	Difference in Diff. Indices	MDT Discr. Index	MC Discr. Index	Difference in Discr. Indices
30	1	0.013	0.214	-0.201	0.000	0.435	-0.435
25	1	0.234	0.250	-0.016	0.095	0.565	-0.470
23	1	0.312	0.226	0.086	-0.095	0.130	-0.225
27	2	0.364	0.714	-0.350	0.571	0.478	0.093
13	2	0.468	0.833	-0.365	0.619	0.304	0.315
14	2	0.610	0.881	-0.271	0.667	0.348	0.319
16	2	0.636	0.643	-0.007	0.333	0.348	-0.015
21	2	0.649	0.571	0.078	0.476	0.304	0.172
22	2	0.662	0.595	0.067	0.190	0.304	-0.114
18	3	0.675	0.714	-0.039	0.571	0.435	0.136
5	3	0.701	0.762	-0.061	0.571	0.391	0.180
15	3	0.701	0.726	-0.025	0.429	0.391	0.038
26	3	0.701	0.690	0.011	0.238	0.609	-0.371
17	3	0.753	0.893	-0.140	0.524	0.261	0.263
24	3	0.805	0.893	-0.088	0.476	0.304	0.172
19	3	0.831	0.869	-0.038	0.333	0.174	0.159
6	3	0.844	0.929	-0.085	0.286	0.087	0.199
8	3	0.844	0.762	0.082	0.381	0.348	0.033
2	3	0.896	0.964	-0.068	0.048	0.130	-0.082
7	3	0.896	0.917	-0.021	0.190	0.217	-0.027
11	3	0.896	0.845	0.051	0.333	0.304	0.029
29	3	0.896	0.821	0.075	0.238	0.174	0.064
3	3	0.909	0.964	-0.055	0.238	0.087	0.151
20	3	0.935	0.845	0.090	0.190	0.304	-0.114
10	3	0.948	0.929	0.019	0.095	0.087	0.008
28	3	0.961	0.940	0.021	0.095	0.043	0.052
9	3	0.974	0.976	-0.002	0.095	0.087	0.008
4	3	0.987	0.988	-0.001	0.048	0.043	0.005
12	3	0.987	0.964	0.023	0.048	0.130	-0.082
1	3	1.000	0.964	0.036	0.000	0.043	-0.043
Average		0.736	0.776	-0.040	0.276	0.262	0.014
Median		0.818	0.845	-0.012	0.238	0.304	0.031

Notes:

(1) "MDT Diff. Category" refers to a ranking of items by their MDT Difficulty Indices, with divisions made at 0.333 and 0.667:

1 = Difficult Item
2 = Intermediate Item
3 = Easy Item.

(2) Other abbreviations are as follows: "MDT" = Multi-digit Answer Bank Format; "MC" = Multiple Choice Format; "Diff." = Item Difficulty; and "Discr." = Item Discrimination.

APPENDIX 3

Table 3: Difficulty and Discrimination Indices for Data Set WG90

Item No.	MDT Diff. Category	MDT Diff. Index	MC Diff. Index	Difference in Diff. Indices	MDT Discr. Index	MC Discr. Index	Difference in Discr. Indices
17	1	0.142	0.198	-0.056	0.068	0.114	-0.046
20	2	0.340	0.407	-0.067	0.659	0.455	0.204
19	2	0.370	0.519	-0.149	0.523	0.500	0.023
5	2	0.382	0.388	-0.006	0.467	0.556	-0.089
2	2	0.442	0.630	-0.188	0.489	0.422	0.067
3	2	0.473	0.576	-0.103	0.689	0.711	-0.022
10	2	0.479	0.485	-0.006	0.556	0.644	-0.088
12	2	0.494	0.574	-0.080	0.614	0.636	-0.022
4	2	0.539	0.600	-0.061	0.556	0.489	0.067
16	2	0.556	0.673	-0.117	0.364	0.364	0.000
1	2	0.600	0.697	-0.097	0.711	0.644	0.067
18	2	0.605	0.735	-0.130	0.659	0.500	0.159
13	2	0.636	0.698	-0.062	0.432	0.386	0.046
14	3	0.704	0.784	-0.080	0.500	0.455	0.045
7	3	0.709	0.739	-0.030	0.489	0.378	0.111
15	3	0.710	0.759	-0.049	0.568	0.523	0.045
8	3	0.721	0.745	-0.024	0.444	0.244	0.200
11	3	0.741	0.759	-0.018	0.500	0.477	0.023
6	3	0.758	0.794	-0.036	0.422	0.422	0.000
9	3	0.776	0.800	-0.024	0.422	0.422	0.000
Average		0.559	0.628	-0.069	0.507	0.467	0.040
Median		0.578	0.685	-0.061	0.500	0.466	0.034

Notes:

(1) "MDT Diff. Category" refers to a ranking of items by their MDT Difficulty Indices, with divisions made at 0.333 and 0.667:

1 = Difficult Item
2 = Intermediate Item
3 = Easy Item.

(2) Other abbreviations are as follows: "MDT" = Multi-digit Answer Bank Format; "MC" = Multiple Choice Format; "Diff." = Item Difficulty; and "Discr." = Item Discrimination.

APPENDIX 3

Table 4: Difficulty and Discrimination Indices for Data Set PG90

Item No.	MDT Diff. Category	MDT Diff. Index	MC Diff. Index	Difference in Diff. Indices	MDT Discr. Index	MC Discr. Index	Difference in Discr. Indices
3	1	0.304	0.630	-0.326	0.500	0.583	-0.083
6	2	0.348	0.565	-0.217	0.500	0.500	0.000
2	2	0.370	0.478	-0.108	0.417	0.500	-0.083
13	2	0.391	0.739	-0.348	0.500	0.583	-0.083
8	2	0.457	0.478	-0.021	0.250	-0.167	0.417
1	2	0.500	0.674	-0.174	0.417	0.333	0.084
15	2	0.500	0.761	-0.261	0.500	0.250	0.250
7	2	0.522	0.804	-0.282	0.833	0.417	0.416
14	2	0.522	0.674	-0.152	0.750	0.583	0.167
9	2	0.543	0.565	-0.022	0.333	0.417	-0.084
11	2	0.565	0.913	-0.348	0.500	0.167	0.333
12	2	0.565	0.826	-0.261	0.667	0.333	0.334
10	2	0.630	0.804	-0.174	0.583	0.250	0.333
Average		0.478	0.685	-0.207	0.519	0.365	0.154
Median		0.500	0.674	-0.217	0.500	0.417	0.167

Notes:

(1) "MDT Diff. Category" refers to a ranking of items by their MDT Difficulty Indices, with divisions made at 0.333 and 0.667:

1 = Difficult Item
2 = Intermediate Item
3 = Easy Item.

(2) Other abbreviations are as follows: "MDT" = Multi-digit Answer Bank Format; "MC" = Multiple Choice Format; "Diff." = Item Difficulty; and "Discr." = Item Discrimination.

APPENDIX 3

Table 5: Difficulty and Discrimination Indices for Data Set MED89

Item No.	MDT Diff. Category	MDT Diff. Index	MC Diff. Index	Difference in Diff. Indices	MDT Discr. Index	MC Discr. Index	Difference in Discr. Indices
36	1	0.130	0.130	0.000	0.060	0.090	-0.030
32	1	0.160	0.160	0.000	0.170	0.110	0.060
39	1	0.170	0.300	-0.130	0.200	0.170	0.030
42	1	0.180	0.450	-0.270	0.350	0.210	0.140
31	1	0.200	0.300	-0.100	0.270	0.470	-0.200
41	1	0.240	0.390	-0.150	0.140	0.060	0.080
16	1	0.250	0.350	-0.100	0.290	0.170	0.120
40	2	0.340	0.740	-0.400	0.280	0.430	-0.150
33	2	0.440	0.710	-0.270	0.360	0.320	0.040
34	2	0.440	0.670	-0.230	0.330	0.260	0.070
35	2	0.480	0.510	-0.030	0.350	0.440	-0.090
38	2	0.490	0.760	-0.270	0.550	0.330	0.220
20	2	0.530	0.680	-0.150	0.460	0.480	-0.020
25	2	0.530	0.810	-0.280	0.230	0.070	0.160
17	2	0.560	0.630	-0.070	0.270	0.180	0.090
21	2	0.610	0.730	-0.120	0.410	0.480	-0.070
23	2	0.630	0.780	-0.150	0.510	0.350	0.160
29	3	0.700	0.940	-0.240	0.540	0.320	0.220
37	3	0.740	0.910	-0.170	0.300	0.370	-0.070
19	3	0.750	0.740	0.010	0.470	0.330	0.140
26	3	0.780	0.870	-0.090	0.190	0.300	-0.110
27	3	0.780	0.880	-0.100	0.460	0.250	0.210
30	3	0.780	0.810	-0.030	0.020	0.120	-0.100
22	3	0.870	0.970	-0.100	0.520	0.370	0.150
28	3	0.870	0.940	-0.070	0.360	0.180	0.180
24	3	0.940	0.970	-0.030	0.400	0.010	0.390
18	3	0.970	1.000	-0.030	0.300		
Average		0.539	0.671	-0.132	0.326	0.264	0.062
Median		0.530	0.740	-0.100	0.330	0.280	0.075

Notes:

(1) "MDT Diff. Category" refers to a ranking of items by their MDT Difficulty Indices, with divisions made at 0.333 and 0.667:

1 = Difficult Item
2 = Intermediate Item
3 = Easy Item.

(2) Other abbreviations are as follows: "MDT" = Multi-digit Answer Bank Format; "MC" = Multiple Choice Format; "Diff." = Item Difficulty; and "Discr." = Item Discrimination.

APPENDIX 4

Table 1. Statistics for Data Set WG84

	Column A	Column B			
Row 1	Source Assessment (Name of Test/Subtest)	WG84			
2	Sample Size (# of students)	MDT = 77 MC = 82			
3	Number of Items (Questions answered)	20			
4	Characteristics of questions	Map place identification			
5	Mean difficulty MDT	0.676			
6	Mean difficulty MC	0.727			
7	Difference = Diff MDT - Diff MC	-0.104			
8	P-Value of Diff P<	0.0079			
9	Correlation (r) and Regression of Difficulty Indices: DiffMC = Intercept + Slope * DiffMDT	r = 0.6521 INTER = 0.4758 SE = 0.0870 SLOPE = 0.4502 SE = 0.1234			
10	P-Value of Correlation of Difficulty P<	P < 0.0019			
11	KR-20 MDT				
12	KR-20 MC				
13	Difference of KR-20's				
14	Mean Discrimination MDT				
15	Mean Discrimination MC				
16	Difference in Mean Discrimination (MDT-MC)				
17	P-Value of Discrimination				
18	Correlation (r) and Regression of Discrimination Indices: DiscrimMC = Intercept + Slope * DiscrimMDT				
19	P-Value of Correlation				
20					
21					

APPENDIX 4

Table 2: Statistics for Data Set WG87

	Column A	Column B	Column C	Column D	
Row 1	Source Assessment (Name of Test/Subtest)	WG87	WG87	WG87	
2	Sample Size (# of students)	161 = 77MDT+84MC			
3	Number of Items (Questions answered)	30	11	16	
4	Characteristics of questions		Mid-difficulty MDT=0.35 to 0.80	Low Diff (easy) MDT=0.80 to 1.00	
5	Mean difficulty MDT	0.736	0.629	0.913	
6	Mean difficulty MC	0.776	0.729	0.911	
7	Difference = Diff MDT - Diff MC	-0.040	-0.100	+0.002	
8	P-Value of Diff P<	0.5541 (Not)	0.0647	0.8684 (not sig)	
9	Correlation (r) and Regression of Difficulty Indices: DiffMC = Intercept + Slope * DiffMDT	r=0.8745 0.20575 SE= 0.06293 0.77460 SE= 0.08118	r=-0.0396 0.7525 SE = 0.1985 -0.0370 SE = 0.3109	r=0.5819 0.3392 SE = 0.2139 0.6258 SE = 0.2338	
10	P-Value of Correlation of Difficulty P<	P < 0.0000	P < 0.9082	0.0182	
11	KR-20 MDT	0.781	0.661	0.765	
12	KR-20 MC	0.780	0.613	0.705	
13	Difference of KR-20's	0.001	0.048	0.060	
14	Mean Discrimination MDT	0.276	0.427	0.193	
15	Mean Discrimination MC	0.262	0.379	0.160	
16	Difference in Mean Discrimination (MDT-MC)	0.014	0.092	0.033	
17	P-Value of Discrim.	P < 0.8474 (Not)	P < 0.1657	P < 0.1854	
18	Correlation (r) and Regression of Discrimination Indices: DiscrimMC = Intercept + Slope * DiscrimMDT	r=0.4715 0.16600 SE= 0.04270 0.34831 SE= 0.12313	r=-0.2293 0.4478 SE = 0.1016 -0.1450 SE = 0.2052	r=0.7325 0.0543 SE = 0.0321 0.5470 SE = 0.1359	
19	P-Value of Correlation	P < 0.0085 (Sig)	P < 0.4985 (not)	P < 0.0013	
20					
21					

BEST COPY AVAILABLE

APPENDIX 4

Table 3: Statistics for Data Set WG90, Part A

	Column A	Column B	Column C	Column D	
Row 1	Source Assessment (Name of Test/Subtest)	WG90	WG90, Test 3	WG90, Test 4	
2	Sample Size (# of students)	160	160	160	
3	Number of Items (Questions answered	20	10	10	
4	Characteristics of questions				
5	Mean difficulty MDT	0.559	0.588	0.611	
6	Mean difficulty MC	0.628	0.645	0.611	
7	Difference = Diff MDT - Diff MC	-0.069	-0.058	-0.081	
8	P-Value of Diff P<	0.0000	0.0115	0.0001	
9	Correlation (r) and Regression of Difficulty Indices: DiffMC = Intercept + Slope*DiffMDT	r = 0.9561 0.1179 SE = 0.0384 0.1928 SE = 0.0660	r = 0.9186 0.1372 SE = 0.0794 0.8644 SE = 0.1314	r = 0.9783 0.1045 SE = 0.0401 0.9553 SE = 0.0715	
10	P-Value of Correlation of Difficulty P<	P < 0.0000	P < 0.0002	0.0000	
11	KR-20 MDT	0.	0.		
12	KR-20 MC	0.	0.		
13	Difference of KR-20's	0.	0.		
14	Mean Discrimination MDT	0.507	0.525	0.489	
15	Mean Discrimination MC	0.467	0.493	0.441	
16	Difference in Mean Discrimination (MDT-MC)	0.0.040	0.031	0.048	
17	P-Value of Discrim.	P < 0.0442	P < 0.2987 (not)	P < 0.0841	
18	Correlation (r) and Regression of Discrimination Indices: DiscrimMC = Intercept + Slope*DiscrimMDT	r = 0.8304 0.0522 SE = 0.0680 0.8190 SE = 0.1253	r = 0.7872 -0.0804 SE = 0.1616 1.0935 SE = 0.3029	r = 0.9060 0.0950 SE = 0.0604 0.7080 SE = 0.1170	
19	P-Value of Correlation	0.0000	0.0072	0.0003	
20					
21					

APPENDIX 4

Table 4: Statistics for Data Set WG90, Part B

	Column A	Column B	Column C		
Row 1	Source Assessment (Name of Test/Subtest)	WG90	WG90		
2	Sample Size (# of students)	160	160		
3	Number of Items (Questions answered	12	7		
4	Characteristics of questions	Mid-diff MDT=0.33 to 0.67	Low Diff (easy) MDT=0.67 to 1.00		
5	Mean difficulty MDT	0.493	0.731		
6	Mean difficulty MC	0.582	0.769		
7	Difference = Diff MDT - Diff MC	-0.089	-0.037		
8	P-Value of Diff P<	0.0001	0.0036		
9	Correlation (r) and Regression of Difficulty Indices: DiffMC = Intercept + Slope*DiffMDT	r = 0.8812 0.0743 SE=0.0876 1.0295 SE=0.1746	r = 0.6677 0.3439 SE = 0.2119 0.5808 SE = 0.2896		
10	P-Value of Correlation of Difficulty P<	P < 0.0002	P < 0.1012 (not)		
11	KR-20 MDT	0.	0.		
12	KR-20 MC	0.	0.		
13	Difference of KR-20's	0.	0.		
14	Mean Discrimination MDT	0.560	0.478		
15	Mean Discrimination MC	0.526	0.417		
16	Difference in Mean Discrimination (MDT-MC)	0.034	0.061		
17	P-Value of Discrim.	P < 0.2052 (not)	P < 0.0680		
18	Correlation (r) and Regression of Discrimination Indices: DiscrimMC = Intercept + Slope*DiscrimMDT	r = 0.6852 0.1310 SE = 0.1350 0.7087 SE = 2370	r = 0.5888 -0.0598 SE = 0.2944 0.9983 SE = 0.6130		
19	P-Value of Correlation	P < 0.0142	P < 0.1643 (not)		
20					
21					

APPENDIX 4

Table 5: Statistics for Data Set PG90

	Column A	Column B	Column C	Column D	Column E
Row 1	Source Assessment (Name of Test)	PG90	PG90	"Harder" items	"Easier" items
2	Sample Size (# of students)	47	42	42	42
3	Number of Items (Questions answered	30 = 15 pairs	26 = 13 pairs; removed 2 low-dif	7 pairs	6 pairs
4	Characteristics of questions	Concepts and terminology	No item has diff < 0.200	MDT dif 0.5 or less	MDT dif > 0.5
5	Mean difficulty MDT	0.44255	0.478	0.410	0.558
6	Mean difficulty MC	0.63262	0.685	0.618	0.764
7	Difference = Diff MDT - Diff MC	-0.19007	-0.207	-0.208	-0.207
8	P-Value of Diff P<	P < 0.0007	P<0.0000 (t-test)	P < 0.0034	P < 0.0072
9	Correlation (r) and Regression of Difficulty Indices: DiffMC = Intercept + Slope*DiffMDT	0.5048 0.31084 SE=0.16094 0.72481 SE=0.34374	0.5953 0.2844 SE = 0.1664 0.8387 SE = 0.3414	0.3118 0.4254 SE = 0.2663 0.4695 SE = 0.6398	0.3662 0.1346 SE = 0.8017 1.1289 SE = 1.4341
10	P-Value of Correlation of Difficulty P<	P < 0.0550	P < 0.0322	P < 0.4960	P < 0.4752
11	KR-20 MDT	0.691	0.682	0.424	0.571
12	KR-20 MC	0.590	0.496	0.374	0.181
13	Difference of KR-20's	0.101	0.186	0.050	0.390
14	Mean Discrimination MDT	0.489	0.519	0.441	0.611
15	Mean Discrimination MC	0.389	0.365	0.369	0.361
16	Difference in Mean Discrimination (MDT-MC)	0.100	0.154	0.072	0.250
17	P-Value of Discrim.	P < 0.1696 (Not)	P < 0.0186	P < 0.3681	P < 0.0203
18	Correlation (r) and Regression of Discrimination Indices: DiscrimMC = Intercept + Slope*DiscrimMDT		0.4255 0.0739 SE = 0.1949 0.5613 SE = 0.3599	0.8479 -0.7084 SE = 0.3068 2.4450 SE = 0.6836	0.3869 0.1697 SE = 0.2362 0.3133 SE = 0.3733
19	P-Value of Correlation		0.1479 (not)	0.0159	0.4486 (not)
20					
21				Note: The discrim. slopes are different at P < 0.0228 level.	

APPENDIX 4

Table 6: Statistics for Data Set MED89, Part A

	Column A	Column B	Column C	Column D	
Row 1	Source Assessment (Name of Test/Subtest)	MED89	MED89	MED89	
2	Sample Size (# of students)	93	93	97	
3	Number of Items (Questions answered	27	9	18	
4	Characteristics of questions		Predefined "Core" questions	Predefined "Difficult"	
5	Mean difficulty MDT	0.539	0.808	0.405	
6	Mean difficulty MC	0.671	0.893	0.561	
7	Difference = Diff MDT - Diff MC	-0.132	-0.086	-0.156	
8	P-Value of Diff P<	P < 0.0000	P < 0.0074	P < 0.0000	
9	Correlation (r) and Regression of Difficulty Indices: DiffMC = Intercept + Slope*DiffMDT	r = 0.9194 0.1842 SE=0.0462 0.9036 SE=0.0773	r = 0.7884 0.3486 SE = 0.1621 0.6744 SE = 0.1989	r = 0.8859 0.1378 SE = 0.0617 1.0438 SE = 0.1367	
10	P-Value of Correlation of Difficulty P<	P < 0.0000	P < 0.0121	P < 0.0000	
11	KR-20 MDT	0.52	0.33	0.41	
12	KR-20 MC	0.46	0.29	0.29	
13	Difference of KR-20's	0.06	0.04	0.12	
14	Mean Discrimination MDT	0.327	0.419	0.286	
15	Mean Discrimination MC	0.264	0.280	0.257	
16	Difference in Mean Discrimination (MDT-MC)	0.062	0.139	0.028	
17	P-Value of Discrim.	P < 0.0308	P < 0.0464	0.3176 (not sig)	
18	Correlation (r) and Regression of Discrimination Indices: DiscrimMC = Intercept + Slope*DiscrimMDT	r = 0.5224 0.0969 SE = 0.0604 0.5123 SE = 0.1698	r = 0.1843 0.1825 SE = 0.2186 0.2328 SE = 0.5069	r = 0.6640 0.0618 SE = 0.0610 0.6845 SE = 0.1927	
19	P-Value of Correlation	P < 0.0060	P < 0.6622 (not)	P < 0.0027	
20					
21					

APPENDIX 4

Table 7: Statistics for Data Set MED89, Part B

	Column A	Column B	Column C	Column D	
Row 1	Source Assessment (Name of Test/Subtest)	MED89	MED89	MED89	
2	Sample Size (# of students)	93	93	93	
3	Number of Items (Questions answered	8	12	7	
4	Characteristics of questions	High Diff (hard) MDT=0.00 to 0.35	Mid-difficulty MDT=0.35 to 0.75	Low Diff (easy) MDT=0.75 to 1.00	
5	Mean difficulty MDT	0.209	0.575	0.856	
6	Mean difficulty MC	0.353	0.739	0.920	
7	Difference = Diff MDT - Diff MC	-0.144	-0.164	-0.064	
8	P-Value of Diff P<	P < 0.0202	P < 0.0001	P < 0.0023	
9	Correlation (r) and Regression of Difficulty Indices: DiffMC = Intercept + Slope*DiffMDT	r = 0.8880 -.1782 SE=0.1170 2.5427 SE=0.5374	r = 0.6308 0.3561 SE = 0.1515 0.6663 SE = 0.2592	r = 0.9071 0.2517 SE = 0.1392 0.7810 SE = 0.1621	
10	P-Value of Correlation of Difficulty P<	P < 0.0032	P < 0.0283	P < 0.0048	
11	KR-20 MDT				
12	KR-20 MC				
13	Difference of KR-20's				
14	Mean Discrimination MDT	0.220	0.398	0.325	
15	Mean Discrimination MC	0.214	0.328	0.205	
16	Difference in Mean Discrimination (MDT-MC)	0.006	0.071	0.120	
17	P-Value of Discrim.	P < 0.8912 (not)	P < 0.0530	P < 0.1888 (not)	
18	Correlation (r) and Regression of Discrimination Indices: DiscrimMC = Intercept + Slope*DiscrimMDT	r = 0.5916 0.0022 SE = 0.1269 0.9618 SE = 0.5351	r = 0.5044 *** 0.1066 SE = 0.1235 0.5546 SE = 0.3002	r = 0.2960 0.1382 SE = 0.1218 0.2056 0.3318	
19	P-Value of Correlation	P < 0.1224	P < 0.0952	P < 0.5690	
20					
21			*** = non-linear curve evident		

BEST COPY AVAILABLE

APPENDIX 4

Table 8: Statistics for Data Set ALLSETS

	Column A	Column B	Column C	Column D	Column E
Row 1	Source Assessment (Name of Test/Subtest)	ALLSETS = Combined All Five Data Sets	ALLSETS	ALLSETS	ALLSETS
2	Sample Size (# of students)				
3	Number of Items (Questions answered)	110 for Difficulty 89 for Discrimination	14 12	46 40	50 37
4	Characteristics of questions		Hardest MDT = 0.00 to 0.33	Mid-difficulty MDT = 0.33 to 0.67	Easiest MDT = 0.67 to 1.00
5	Mean difficulty MDT	0.614	0.189	0.519	0.820
6	Mean difficulty MC	0.713	0.339	0.675	0.854
7	Difference = Diff MDT - Diff MC	-0.099	-0.150	-0.155	-0.034
8	P-Value of Diff P<	P < 0.0000	P < 0.0041	P < 0.0000	P < 0.0009
9	Correlation (r) and Regression of Difficulty Indices: DiffMC = Intercept + Slope*DiffMDT	r = 0.8501 0.2547 SE=0.0293 0.7470 SE=0.0445	r = 0.3916 0.1867 SE = 0.1124 0.8070 SE = 0.5474	r = 0.4386 0.3710 SE = 0.0953 0.5845 SE = 0.1805	r = 0.6936 0.2409 SE = 0.0925 0.7475 SE = 0.1121
10	P-Value of Correlation of Difficulty P<	P < 0.0000	P < 0.1668 (not)	P < 0.0023	P < 0.0000
11	KR-20 MDT	0.	0.		
12	KR-20 MC	0.	0.		
13	Difference of KR-20's	0.	0.		
14	Mean Discrimination MDT	0.378	0.171	0.489	0.325
15	Mean Discrimination MC	0.324	0.259	0.397	0.265
16	Difference in Mean Discrimination (MDT-MC)	0.054	-0.088	0.092	0.060
17	P-Value of Discrim.	P < 0.0027	P < 0.2721 (not)	P < 0.0026	P < 0.0062
18	Correlation (r) and Regression of Discrimination Indices: DiscrimMC = Intercept + Slope*DiscrimMDT	r = 0.6189 0.1160 SE = 0.0319 0.5498 SE = 0.0748	r = 0.3594 0.1857 SE = 0.0817 0.4292 SE = 0.3524	r = 0.5213 0.1184 SE = 0.0776 0.5702 SE = 0.1514	r = 0.6709 0.0794 SE = 0.0396 0.5721 SE = 0.1069
19	P-Value of Correlation	P < 0.0000	P < 0.2521 (not)	P < 0.0006	P < 0.0000
20					
21					

Notes:

(1) For the discrimination calculations, N = 89 pairs of items because WG84 (with 20 items) and MED89 (one item) do not have calculated discrimination values. Remember that MED89 used a different discrimination calculation (point-biserial).

APPENDIX 4

Table 9: Statistics for Data Set GEOG3, Part A

	Column A	Column B	Column C	Column D	Column E
Row 1	Source Assessment (Name of Test/Subtest)	GEOG3	GEOG3	GEOG3	GEOG3
2	Sample Size (# of students)				
3	Number of Items (Questions answered	63	5	30	28
4	Characteristics of questions		Hardest MDT=0.00 to 0.33	Mid-difficulty MDT=0.33 to 0.67	Easiest MDT= 0.67 to 1.00
5	Mean difficulty MDT	0.627	0.201	0.507	0.831
6	Mean difficulty MC	0.710	0.304	0.650	0.848
7	Difference = Diff MDT - Diff MC	-0.084	-0.103	-0.143	-0.017
8	P-Value of Diff P<	P < 0.0000	P < 0.2249 (not)	P < 0.0000	P < 0.0856
9	Correlation (r) and Regression of Difficulty Indices: DiffMC = Intercept + Slope*DiffMDT	r = 0.8512 0.2475 SE = 0.0388 0.7385 SE = 0.0583	r = 0.5035 0.1555 SE = 0.1680 0.7369 SE = 0.7301	r = 0.4703 0.3207 SE = 0.1189 0.6492 SE = 0.2303	r = 0.8650 0.2217 SE = 0.0708 0.7536 SE = 0.0857
10	P-Value of Correlation of Difficulty P<	P < 0.0000	P < 0.3871 (not)	P < 0.0088	P < 0.0000
11	KR-20 MDT	0.	0.		
12	KR-20 MC	0.	0.		
13	Difference of KR-20's	0.	0.		
14	Mean Discrimination MDT	0.399	0.114	0.528	0.313
15	Mean Discrimination MC	0.349	0.365	0.419	0.270
16	Difference in Mean Discrimination (MDT-MC)	0.050	-0.252	0.109	0.043
17	P-Value of Discrim.	P < 0.0129	P < 0.0450	P < 0.0049	P < 0.0230
18	Correlation (r) and Regression of Discrimination Indices: DiscrimMC = Intercept + Slope*DiscrimMDT	r = 0.6226 0.1345 SE = 0.0391 0.5359 SE = 0.0862	r = 0.6353 0.2928 SE = 0.1078 0.6391 SE = 0.4485	r = 0.4708 .4798? 0.1218 SE = 0.1064 0.5626 SE = 0.1944	r = 0.7525 0.0614 SE = 0.0415 0.6669 SE = 0.1145
19	P-Value of Correlation	P < 0.0000	P < 0.2494 (not)	P < 0.0073	P < 0.0000
20					
21					

APPENDIX 4

Table 10: Statistics for Data Set GEOG3, Part B

	Column A	Column B	Column C		
Row 1	Source Assessment (Name of Test/Subtest)	GEOG3	GEOG3		
2	Sample Size (# of students)				
3	Number of Items (Questions answered)	46	12		
4	Characteristics of questions	Mid-difficulty MDT=0.33 to 0.90	Very Easy MDT=0.90 to 1.00		
5	Mean difficulty MDT	0.591	0.940		
6	Mean difficulty MC	0.698	0.926		
7	Difference = Diff MDT - Diff MC	-0.107	+0.014		
8	P-Value of Diff P<	P < 0.0000	P < 0.3078 (not)		
9	Correlation (r) and Regression of Difficulty Indices: DiffMC = Intercept + Slope*DiffMDT	r = 0.6588 0.3348 SE = 0.0644 0.6148 SE = 0.1059	r = 0.5833 0.1534 SE = 0.3407 0.8220 SE = 0.3619		
10	P-Value of Correlation of Difficulty P<	P < 0.0000	P < 0.0470		
11	KR-20 MDT	0.	0.		
12	KR-20 MC	0.	0.		
13	Difference of KR-20's	0.	0.		
14	Mean Discrimination MDT	0.500	0.135		
15	Mean Discrimination MC	0.402	0.137		
16	Difference in Mean Discrimination (MDT-MC)	0.098	-0.003		
17	P-Value of Discrim.	P < 0.0001	P < 0.9374 (not)		
18	Correlation (r) and Regression of Discrimination Indices: DiscrimMC = Intercept + Slope*DiscrimMDT	r = 0.4475 0.1408 SE = 0.0815 0.5223 SE = 0.1574	r = 0.7180 0.0471 SE = 0.0341 0.6698 SE = 0.2053		
19	P-Value of Correlation	P < 0.0018	P < 0.0088		
20					
21					